Order-Preserving Nonparametric Regression, With Applications to Conditional Distribution and Quantile Function Estimation

Peter HALL and Hans-Georg MÜLLER

In some regression problems we observe a "response" Y_{ti} to level t of a "treatment" applied to an individual with level X_i of a given characteristic, where it has been established that response is monotone increasing in the level of the treatment. A related problem arises when estimating conditional distributions, where the raw data are typically independent and identically distributed pairs (X_i, Z_i), and Y_{ti} denotes the proportion of Z_i's that do not exceed t. We expect the regression means $g_t(x) = E(Y_{ti}|X_i = x)$ to enjoy the same order relation as the responses, that is, $g_t \le g_s$ whenever $s \le t$. This requirement is necessary to obtain bona fide conditional distribution functions, for example. If we estimate g_t by passing a linear smoother through each dataset $\mathcal{X}_t = \{(X_i, Y_{ti}) : 1 \le i \le n\}$, then the order-preserving property is guaranteed if and only if the smoother has nonnegative weights. However, in such cases the estimators generally have high levels of boundary bias. On the other hand, the order-preserving property usually fails for linear estimators with low boundary bias, such as local linear estimators, or kernel estimators employing boundary kernels. This failure is generally most serious at boundaries of the distribution of the explanatory variables, and ironically it is often in just those places that estimation is of greatest interest, because responses there imply constraints on the larger population. In this article we suggest nonlinear, order-invariant estimators for nonparametric regression, and discuss their properties. The resulting estimators are applied to the estimation of conditional distribution functions at endpoints and also changepoints. The availability of bona fide distribution function estimators at endpoints also enables the computation of changepoint diagnostics that are based on differences in a suitable norm between two estimated conditional distribution functions, obtained from data that fall into one-sided bins.

KEY WORDS: Bias reduction; Boundary effect; Changepoint; Endpoint; Linear methods; Local linear estimator; Monotonicity; Nadaraya–Watson estimator; Prediction.

1. INTRODUCTION

1.1 Regression Problems Requiring Order-Preserving Solutions

Suppose we observe a "response," Y_{ti}, to level t of a "treatment" applied to an individual with level X_i of a given characteristic, where it has been established that response is monotone across the range of the treatment. For instance, Y_{ti} might represent the height at age t years of the ith child in a growth study, where the child's mother's height was X_i units. We expect the regression means, $E(Y_{ti}|X_i = x)$, to be monotone in t for fixed x, but monotonicity of the estimators is not assured by conventional nonparametric regression smoothers. Difficulties with monotonicity can be particularly acute toward the ends of the design interval. In many applications it is on just those places that the majority of interest centers.

A major motivating example to consider in this setting is the nonparametric estimation of a conditional distribution function. This problem is of interest in its own right, but also has applications to the estimation of conditional density functions and quantile functions, which usually are derived from a suitable conditional distribution function estimate. Suppose we observe a sequence of independent and identically distributed data pairs (X_i, Z_i), for $1 \le i \le n$, and wish to construct an estimator $\widehat{F}(\cdot|x)$ of the conditional distribution function $F(t|x) \equiv P(Z \le t|X = x)$. This problem admits a simple solution in terms of nonparametric regression, because if we define $Y_{ti} = I(Z_i \le t)$, then $F(t|x)$ equals the mean of Y_{ti} given $X_i = x$. We would, of course, require $\widehat{F}(t|x)$ to be a bona fide distribution function estimate, that is, to be monotone increasing in t, but this property is not guaranteed by passing conventional regression smoothers through the dataset $\mathcal{X}_t = \{(X_i, Y_{ti}) : 1 \le i \le n\}$.

Estimating a conditional distribution function at a boundary or endpoint of the support of the covariates is of special interest for two reasons. First, we may wish to construct prediction intervals for a new observation that will be made right at the boundary of the current domain of the covariate, as is often the case when observations are made sequentially involving regular small increments of the covariate, such as in quality control or environmental applications. Such prediction intervals are conveniently based on estimated conditional distributions. Second, for the detection and estimation of changepoints that may involve changes in features of the conditional distribution that are more general than just mean changes, the estimation of bona fide conditional distribution functions at endpoints will provide an essential tool. Differences in a suitable metric between estimated left- and right-sided conditional distribution functions, based on one-sided windows placed around an assumed changepoint location and taken as a function of this assumed location, provide changepoint diagnostics.

This article addresses this problem by introducing the more general perspective of *order-preserving nonparametric regression*. Specifically, we address sequences of datasets $\mathcal{X}_t = \{(X_i, Y_{ti}) : 1 \le i \le n\}$, for $t \in \mathcal{T}$, where the explanatory variables X_i are common to each $\mathcal{X}_t$, and $\mathcal{T}$ denotes an interval that might be either bounded or unbounded, either discrete or in the continuum. The pairs (X_i, Y_{ti}) may often be regarded as observations of a generic (X, Y_t), say. The Y_{ti}'s are ordered, in the sense that, for each i, $Y_{si} \le Y_{ti}$ whenever $s \le t$. Therefore, we expect the regression means $g_t(x) = E(Y_t|X = x)$ to

Peter Hall is Professor, Centre for Mathematics and Its Applications, Australian National University, Canberra, ACT 0200, Australia. Hans-Georg Müller is Professor, Department of Statistics, University of California, Davis, CA 95616 (E-mail: *mueller@wald.ucdavis.edu*). The research of H.G.M. was partially supported by National Science Foundation grants DMS-9971602 and DMS-0204869 and a Visiting Fellowship to the Centre for Mathematics and Its Applications, Australian National University. The authors thank two referees and an associate editor for most helpful and constructive comments.

Journal of the American Statistical Association
September 2003, Vol. 98, No. 463, Theory and Methods
DOI 10.1198/016214503000000512

be ordered: $g_s(x) \le g_t(x)$ whenever $s \le t$ and x is in the support interval $\mathcal{I}$ of the distribution of X. We wish to construct a sequence of estimators $\{\hat{g}_t : t \in \mathcal{T}\}$ of the set of functions $\{g_t : t \in \mathcal{T}\}$, which enjoys the same ordering property. If it does, then we say the estimators are *order preserving* on $\mathcal{I}$. We shall suggest order-preserving regression smoothers with relatively low bias, particularly at the extremities of the design interval, and describe their properties.

1.2 Existing Order-Preserving Methods

Quotient methods for nonparametric regression, for example, the Nadaraya–Watson estimator, are order preserving if based on nonnegative kernel weights. This follows from the fact that such techniques (a) are linear in the response variables Y_i and (b) have the *positivity property*; that is, whenever the response variables are all nonnegative, the estimator itself is nonnegative. More generally, a linear estimator is order preserving if and only if it has the positivity property.

Several recent examples of quotient methods are based on the Nadaraya–Watson estimator. They include the identity-reproducing regression or mass-centered smoothing techniques, discussed by Müller and Song (1993) and Mammen and Marron (1997); the biased bootstrap form of the Nadaraya–Watson estimator (Hall and Presnell 1999); and some, although not all, data-sharpening techniques (Choi, Hall, and Rousson 2000). However, all these methods suffer excessive bias at the boundaries. Specifically, although they have $O(h^2)$ bias in the interior, where h denotes bandwidth, this rate deteriorates to $O(h)$ near a boundary.

Convolution-typeestimators, such as those of Gasser–Müller and Priestley–Chao type (see, for example, Wand and Jones 1995, p. 130ff; Simonoff 1996, p. 138), are also order preserving, as long as positive kernel weights are used, because they are linear and have the positivity property. For both convolution-type and quotient-type estimators, the positivity property holds if the kernels used are nonnegative. On the other hand, local linear methods, which are well known for their high level of resistance to boundary effects (see, for example, Fan 1992; Fan and Gijbels 1992), lack the positivity property and are not order preserving, even while employing a nonnegative kernel or weight function. Moreover, they suffer this deficiency even in the asymptotic limit—in an important class of problems the probability that a local linear estimator, computed for a particular realization, is not order preserving at the boundary converges to 1 as sample size increases; see Section 3.1.

In fact, no linear, kernel-type estimator that enjoys the positivity property can have better than $O(h)$ bias at the boundary, where h denotes the estimator's bandwidth; see Section 3.1. Equivalently, no order-preserving kernel-type estimator with better than $O(h)$ boundary bias can be linear. Therefore, nonlinear estimators must be used if we are to obtain an order-preserving estimator with $O(h^2)$ bias across the full design interval. In particular, none of the estimators discussed previously is suitable. This also includes traditional methods for alleviating edge effect problems based on boundary kernels, as it can be easily shown that suitable boundary kernel functions cannot be restricted to be nonnegative.

1.3 The Relevance of Conditional Distribution Function Estimation at Boundaries

As mentioned previously, a major motivation for order-preserving regression is the desire to obtain bona fide distribution function estimates (see, for example, Hall, Wolff, and Yao 1999; Peracchi 2002, for some recent work on this problem). Such estimates are important for a variety of purposes, one of which is estimation of conditional density functions that are implicitly derived via conditional distribution functions. Another application is the nonparametric estimation of conditional quantile functions as inverses of conditional distribution functions, a problem that has been studied by Bhattacharya and Gangopadhyay (1990) and Yu and Jones (1998), among others. There is particular motivation for estimating bona fide conditional distribution functions exactly at those covariate levels where the problem is hardest, namely, at or near endpoints and changepoints defined in terms of the covariate level.

2. METHODOLOGY

Although it is clear from the discussion in Section 1 that order-preserving methods with good boundary bias properties are necessarily nonlinear, the linearity of techniques such as those of Nadaraya and Watson, Gasser and Müller, and Priestley and Chao is partly responsible for their order-preserving property. Therefore, we seek a method that combines the best of both worlds, that is, that corrects for boundary bias without losing the important features of linearity. This leads us to suggest that nonlinear methods be used to impute ordered "pseudo-data" on the sides of boundaries away from the real dataset and that then relatively conventional linear methods be applied to the new, larger dataset, to produce an estimator that is order preserving.

Data imputation by reflection in the boundaries, much as discussed by Schuster (1985), Silverman (1986, p. 30f), and Cline and Hart (1991) in the context of density estimation, leads directly to an order-preserving estimator. It enjoys only $O(h)$ bias at the boundaries, however. Hall and Wehrly (1991) suggested an alternative data imputation method, which involves reflection in points on the boundaries, but although it has good boundary bias properties it fails to be order preserving. We propose an order-preserving version of the Hall–Wehrly technique, which attains the desirable $O(h^2)$ boundary bias rate, as follows.

Assume we have an ordered sequence of datasets $\mathcal{X}_t$, as suggested in Section 1.1; in particular, $Y_{si} \le Y_{ti}$ whenever $s \le t$. Suppose the distribution of the explanatory variables X_i is supported on an interval $\mathcal{I} = [a, b]$, which we call the design interval. Let $\hat{g}_{\mathrm{LL},t}$, for $t \in \mathcal{T}$, denote a local linear estimator of g_t computed from data in $\mathcal{X}_t$. It is defined by minimizing the weighted sum of squares

$$\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)\left[Y_{ti} - \{\beta_0 + \beta_1(x - X_i)\}\right]^2,$$

with respect to β_0, β_1 and setting $\hat{g}_{\mathrm{LL},t} = \hat{\beta}_0$. Here K is a kernel function and h the sequence of bandwidths.

For $x = a$ or $x = b$ consider the sequence $\mathcal{U}(x) = \{\hat{g}_{\mathrm{LL},t}(x) : t \in \mathcal{T}\}$. Ideally, the elements of each $\mathcal{U}(x)$ would be monotone increasing in t, but this is unlikely to be the case for the endpoints $x = a, b$. We wish to "monotonize" $\hat{g}_{\mathrm{LL},t}(x)$ at just

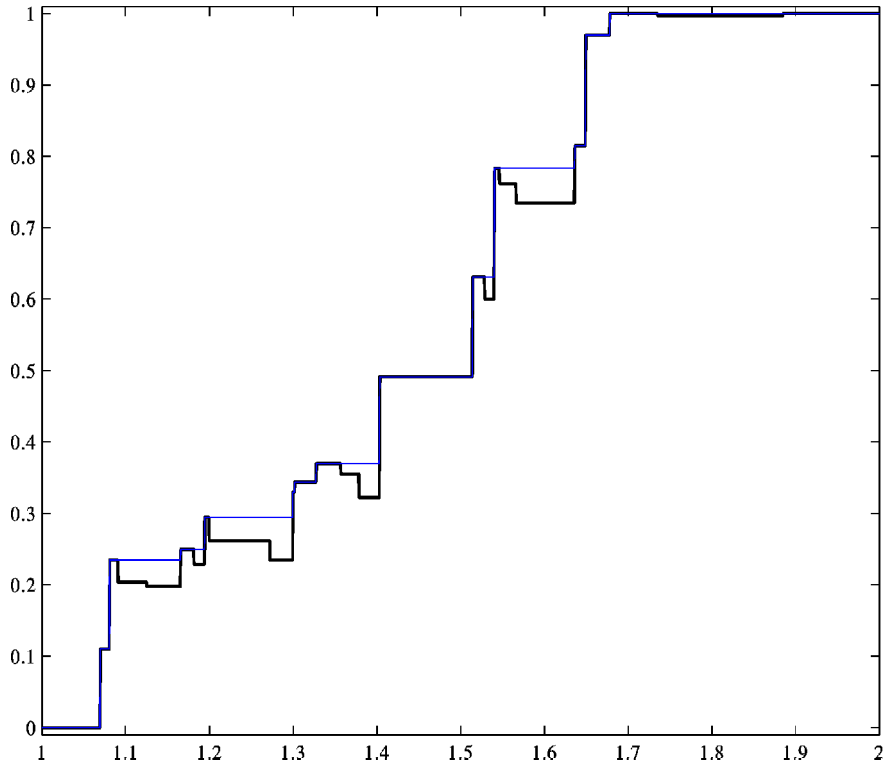

Figure 1. Filling in the "Valleys" of the Initial Distribution Function Estimate $\widehat{F}_{LL}(\cdot\,|b)$ Obtained by Local Linear Fitting at a Right Endpoint b (solid line) Results in a Monotone Increasing Estimate $\widetilde{F}(t|b)$ (thin line) as t Increases (simulated example with $n = 100$ data).

those places. There are several ways of achieving this end; we consider one particular method later, where we denote the monotonized version of $\hat{g}_{\mathrm{LL},t}(x)$ (monotone in the sense of a function of t for fixed x) by $\tilde{g}_t(x)$. In effect, it fills in the valleys of $\hat{g}_{\mathrm{LL},t}(x)$ by horizontal lines. For the special case of estimating a conditional distribution function, this is indicated in Figure 1 for an example dataset. The initial nonmonotone (in t) estimate $\widehat{F}_{\mathrm{LL}}(\cdot, x)$ of the conditional distribution function is monotonized to produce the version $\widehat{F}(\cdot|x)$.

For simplicity, we assume that $\mathcal{T}$ is closed and bounded, although our methods lead to a more general definition of $\tilde{g}_t$. The prescription given later is descriptive, but can readily be given in mathematically rigorous terms. Let $\mathcal{V}(x)$ be the set of points t such that $\hat{g}_{\mathrm{LL},u}(x) \le \hat{g}_{\mathrm{LL},t}(x)$ for all $u \le t$ and also such that, on passing to a point in $\mathcal{T}$ immediately to the right of t, $\hat{g}_{\mathrm{LL},t}(x)$ turns strictly downward, rather than taking a nondecreasing path. In practical applications $\mathcal{V}(x)$ would be finite, so we write it in strictly increasing order as $\mathcal{V}(x) = \{s_1, \ldots, s_N\}$. For each $s \in \mathcal{V}$, let $u(s)$ be the point $t \in \mathcal{T}$ at which $\hat{g}_{\mathrm{LL},t}(x)$ first recovers the same level as, or a greater level than, $\hat{g}_{\mathrm{LL},s}(x)$. Put $t_i = u(s_i)$ and define $\tilde{g}_t(x) = \hat{g}_{\mathrm{LL},s_i}(x)$ if $s_i \le t < t_i$ for some i, and $\tilde{g}_t(x) = \hat{g}_{\mathrm{LL},t}(x)$ otherwise. (Definitions near the boundary are handled in the obvious way; see Fig. 1.) Then $\tilde{g}_t(x)$ is nondecreasing in $t \in \mathcal{T}$.

For each $t \in \mathcal{T}$, we then compute pseudo-data by projecting the points in $\mathcal{X}_t$ through both $\tilde{g}_t(a)$ and $\tilde{g}_t(b)$; see Figure 2 for the distribution function case. This produces a new dataset $\mathcal{X}'_t = \{(X'_i, Y'_{ti}) : 1 \le i \le 3n\}$, say, where, without loss of generality, $X_1 \le \cdots \le X_{3n}$. Thus, the middle n pairs (X'_i, Y'_{ti}) are the original data, the first n are pseudo-data on the left side of the lower boundary, and the last n are pseudo-data on the right side of the upper boundary.

We compute $\hat{g}_t$ by fitting a kernel-type linear estimator through $\mathcal{X}'_t$, assuming that the estimator has the form

$$\hat{g}_t(x) = \sum_{i=1}^{3n} w_i(x) Y'_{ti}, \tag{2.1}$$

where, for a constant $C > 0$, the weights satisfy:

$$\text{for } 1 \le i \le 3n, \quad w_i(\cdot) \ge 0 \quad \text{and} \tag{2.2}$$

$$w_i(x) = 0 \quad \text{whenever } |x - X'_i| > Ch;$$

$$\text{for } 1 \le i \le n \text{ and } 0 \le x \le Ch, \quad w_{n-i+1}(a + x) \le w_{n+i}(a + x) \quad \text{and} \tag{2.3}$$

$$w_{2n-i+1}(b - x) \le w_{2n+i}(b - x).$$

Condition (2.2) is, of course, satisfied by kernel-type estimators based on nonnegative kernels supported in the interval $[-C, C]$. Condition (2.3) is also typically satisfied. To appreciate why, observe that, by definition of our reflection method, the set of design points of the pseudo-data generated on the left side of a (respectively, on the right side of b) is the reflection in $x = a$ (respectively, in $x = b$) of the set of design points of the real data. When $\hat{g}_t$ is a Nadaraya–Watson estimator, the choice we make in the illustrating examples, the weight $w_i(x)$ equals the ratio of a single kernel weight $K_i(x) = K\{(x - X_i)/h\}$ to the sum $\sum_j K_j(x)$. If K is symmetric, unimodal, and supported on $[-C, C]$, then, provided $h \le (b - a)/(2C)$, this construction implies that at $a + x$, with $0 \le x \le Ch$, the denominators of

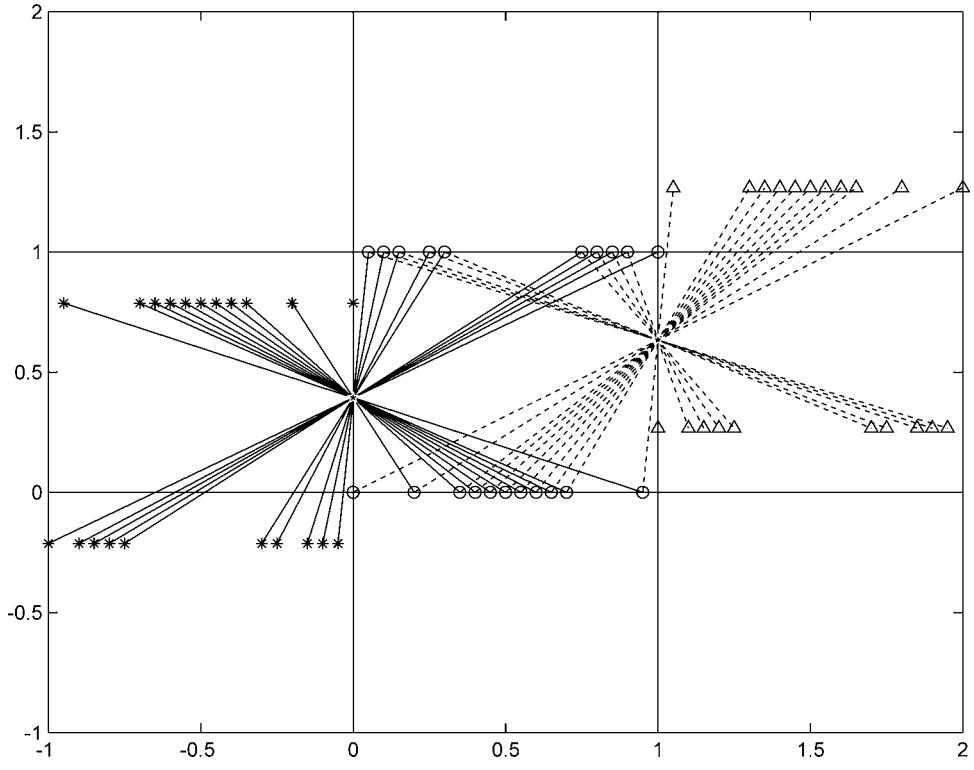

Figure 2. Generation of Pseudo-Data $\mathcal{X}_t'$ by Reflecting Original Data $\mathcal{X}_t$ (confined to domain [a, b] = [0, 1]) at the Points (0, $\widetilde{F}(t|0)$) and (1, $\widetilde{F}(t|1)$). The data (X_i, Y_{ti}) illustrated here were simulated, using sample size $n = 20$ and a single value for t. The original data (○) are indicator variables $Y_{ti} = I(Z_i \le t)$, as used in the application for estimating a conditional distribution function. They are augmented by pseudo-data to the left (∗) and to the right (△).

$w_{n-i+1}(\cdot)$ and $w_{n+i}(\cdot)$ are identical, but (by virtue of the unimodality) the numerator of the latter is not less than that of the former. This property, along with its counterpart for the other boundary, implies (2.3).

This last smoothing step, when implemented with a Nadaraya–Watson quotient-type kernel estimator, is illustrated for the conditional distribution function case in Figure 3 for an example dataset. The resulting distribution function estimator is denoted by $\widehat{F}(\cdot|x)$.

The following result, proved in Section 3, demonstrates that $\hat{g}_t$ has the required properties.

Theorem 2.1. If $\hat{g}_t$ is defined by (2.1), if the weights satisfy (2.2) and (2.3), and if $0 < h \le \frac{1}{2}(b-a)$, then $\hat{g}_t$ is order preserving on $\mathcal{I}$.

3. THEORETICAL PROPERTIES

3.1 Problems Suffered by Linear Estimators

Let $\hat{g}(x) = \sum_i w_i(x) Y_i$ be a linear estimator of $g(x) = E(Y|X = x)$, computed from the independent and identically distributed data $\mathcal{X} = \{(X_i, Y_i) : 1 \le i \le n\}$. We shall say that $\hat{g}$ is of kernel type with bandwidth h if the w_i's are functionals of $X_1, \ldots, X_n$ satisfying, for constants $C_1, C_2, C_3 > 0$, for all x in the design interval, and for all sufficiently large n,

$$w_i(x) = 0 \quad \text{whenever } |x - X_i| > C_1 h, \tag{3.1}$$

and

$$\sum_{i:|x-X_i|>C_2 h} w_i(x) \ge C_3.$$

Conventional kernel estimators, such as those of Nadaraya–Watson, Gasser–Müller, or Priestley–Chao type, satisfy this condition with probability 1 when the kernel is nonnegative and compactly supported, when the design density is bounded away from 0 on the design interval, and when the bandwidth satisfies the mild conditions $h = h(n) \to 0$ and $nh/(\log n)^{1/2} \to \infty$.

Assume the design interval is $[a, b]$. Our first result shows, in effect, that no kernel-type linear estimator with the positivity property can have better than $O(h)$ bias at the boundary. Because a linear estimator is order preserving if and only if it has the positivity property, then linear, order-preserving estimators fail to have good bias properties.

Theorem 3.1. If $\hat{g}$ is a kernel-type linear estimator with the positivity property, and if in the case where $g(x) \equiv C$ (a constant) we have

$$E\{\hat{g}(a)|X_1, \ldots, X_n\} = C + o_p(h) \tag{3.2}$$

as $h \to 0$, then whenever g has a continuous derivative on $[a, b]$ and $g'(a) \ne 0$, there exists $\epsilon > 0$ such that the probability that

$$\left|E\{\hat{g}(a)|X_1, \ldots, X_n\} - g(a)\right| > \epsilon h$$

converges to 1 as $n \to \infty$.

In general, local linear estimators fail to have the positivity property, although it might be thought that this is only a rare defect—for large samples local linear estimators of regression means might be expected to be order-preserving except in unusual cases. Unfortunately, this is not true. We shall show that, with probability tending to 1, local linear methods fail to be

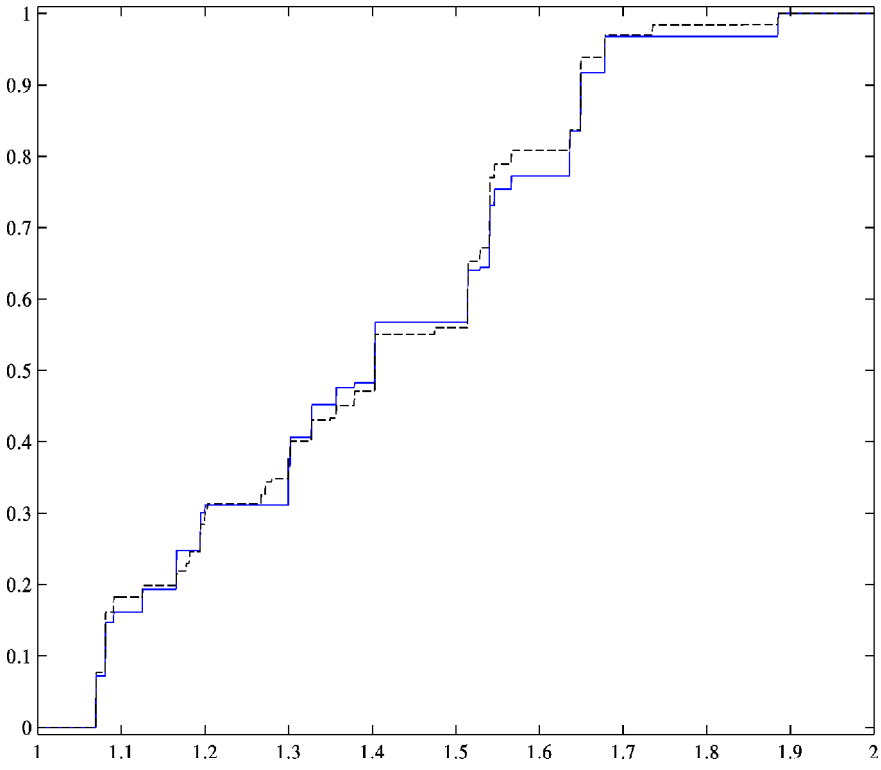

Figure 3. Comparison of Monotonized Intermediate Conditional Distribution Function Estimate $\widetilde{F}(\cdot, x)$ (solid) and the Final Estimate $\widehat{F}(\cdot, x)$ (dashed), Constructed From the Pseudo-data $\mathcal{X}_t'$ Obtained in the Reflection Step. Here, for simulated data with $n = 100$, h = .2 when estimating at the point $x = .95$ near the right endpoint at b = 1.

order preserving in the important case of distribution function estimation. This is only one member of a large class of examples for which local linear methods fail to be order preserving.

The problems arise because of the way local linear methods deal with edge effects. On the other hand, if the design density is supported on a compact interval $\mathcal{I}$ and bounded away from 0 there, then, with probability tending to 1 as $n \to \infty$, local linear estimators have the positivity property when applied to data pairs whose design component is confined to a compact subinterval of $\mathcal{I}$ that does not include the endpoints of $\mathcal{I}$. Therefore, local linear estimators seldom fail to be order preserving in the interior of $\mathcal{I}$. We compile some assumptions as follows.

C0. Let (X_i, Z_i), for $1 \le i \le n$, be a sequence of independent and identically distributed random 2-vectors, with the marginal distribution of X being supported on an interval $\mathcal{I} = [a, b]$, and the marginal density continuous and nonvanishing there. Construct an estimator $\widehat{F}_{\mathrm{LL}}(t|x)$ of $F(t|x) = P(Z \le t | X = x)$ by passing a local linear smoother through the pairs (X_i, Y_{ti}), where $Y_{ti} = I(Z_i \le t)$, using a bounded, compactly supported, symmetric, piecewise continuous, nonnegative kernel K and a bandwidth h. Assume that $h = h(n) \to 0$ and $nh \to \infty$ as $n \to \infty$, which are minimum conditions for weak consistency. We also require that, for $x = a$ and b, $F(\cdot|x)$ be nonsingular.

Theorem 3.2. Under conditions C0 and with probability tending to 1 as $n \to \infty$, there exist intervals $[a, \hat{x}_1]$ and $[\hat{x}_2, b]$, with $\hat{x}_1, \hat{x}_2$ stochastic and $a < \hat{x}_1 \le \hat{x}_2 < b$, such that whenever x is an element of either interval, $\widehat{F}_{\mathrm{LL}}(\cdot|x)$ is not monotone nondecreasing.

It may be proved, under slightly more restrictive conditions, that the lengths of the intervals $[a, \hat{x}_1]$ and $[\hat{x}_2, b]$ are both $O_p(h)$ and that they may be chosen so that, with probability tending to 1, their lengths exceed ϵh, provided $\epsilon > 0$ is taken sufficiently small (but fixed). Furthermore, the problems evinced by Theorem 3.2 are not overcome by modifying local linear estimators in conventional ways. For example, incorporating a ridge parameter does not alleviate the difficulties, because it alters only the denominators of the smoothing weights used to construct local linear estimators; the nonpositivity of local linear methods, which is the root cause of the problems, is caused by the numerators of the smoothing weights.

3.2 Properties of the Order-Preserving Estimator $\hat{g}_t$

For the sake of definiteness, we shall take $\hat{g}_t$, defined in Section 2 and computed from the data and pseudo-data, to be a standard Nadaraya–Watson estimator, although our results apply to other estimator types as well. The Nadaraya–Watson kernel estimator for $g_t(x)$, obtained from the pseudo-data (X_i', Y_{ti}'), is

$$\hat{g}_t(x) = \sum_{i=1}^{3n} K\left(\frac{x - X_i'}{h}\right) Y_{ti}' \Big/ \sum_{i=1}^{3n} K\left(\frac{x - X_i'}{h}\right).$$

With the aim of obtaining useful upper bounds, we compile the following assumptions:

C1. For each t the 2-vectors (X_i, Y_{ti}), for $1 \le i < \infty$, are independent and identically distributed; the common distribution of the independent random variables X_i is continuous on a compact interval $\mathcal{I}$, has a density that is bounded away from 0 and has two bounded derivatives

there, and vanishes off $\mathcal{I}$; for some $\epsilon > 0$,

$$\sup_{t\in\mathcal{T},|u|\le\epsilon} E\Big(\exp\big[u\{Y_{ti} - E(Y_{ti}|X_i)\}\big]\Big) < \infty; \quad (3.3)$$

the regression mean functions $g_t(x) = E(Y_{ti}|X_i = x)$, for $t \in \mathcal{T}$, and their first two derivatives, are bounded uniformly with respect to both $x \in \mathcal{I}$ and $t \in \mathcal{T}$; the bandwidths h_1 used to construct the local linear estimator $\hat{g}_{\mathrm{LL},t}$ from the dataset $\mathcal{X}_t$, and h_2 used to construct the Nadaraya–Watson estimator $\hat{g}_t$ from $\mathcal{X}_t'$, both satisfy $n^\epsilon h_j \to 0$ and $n^{1-\epsilon}h_j \to \infty$ as $n \to \infty$ for some $\epsilon > 0$; the kernels used for either estimator are symmetric, compactly supported, nonnegative, and Hölder continuous on the real line; the kernel used for $\hat{g}_t$ is unimodal; and the number of elements of $\mathcal{T} = \mathcal{T}(n)$ increases no more than polynomially fast in n.

Theorem 3.3. Assume conditions C1. Then, with probability 1,

$$\sup_{t\in\mathcal{T}}\sup_{x\in\mathcal{X}}\big|\hat{g}_t(x) - g_t(x)\big| = O\big\{(nh)^{-1/2}(\log n)^{1/2} + h^2\big\}. \quad (3.4)$$

The rate of convergence asserted at (3.4) is the best possible for even a single t and for distributions satisfying conditions C1. In fact, it may be proved that in the case of a single regression, and under additional regularity conditions,

$$\sup_{x\in\mathcal{X}}\big|\hat{g}_t(x) - E\big\{\hat{g}_t(x)|X_1,\dots,X_n\big\}\big| = C\big\{1 + o(1)\big\}(nh)^{-1/2}(\log n)^{1/2},$$

with probability 1, where $C > 0$ is a constant; and $E\{\hat{g}_t(x)|X_1, \dots, X_n\} = h^2\gamma_t(x) + o(h^2)$, with probability 1, where γ_t is a nonvanishing function. Therefore, the convergence rate at (3.4) is also best possible when it is asserted uniformly in t. We note that the final assumption in conditions C1, about the rate at which the number of elements (size) of $\mathcal{T}(n)$ increases, is usually adequate even when $\mathcal{T}$ is infinite, and also that there is no difficulty extending our methods and results to the fixed design case where $X_1, \dots, X_n$ are nonstochastic and spaced according to a smooth positive design density.

3.3 Application to Distribution Function Estimation

Depending on the model that generates the ordered datasets $\mathcal{X}_t$, alternative methods can be used to derive rates of convergence of $\hat{g}_t$ to g_t in integral metrics, not requiring the logarithmic factor on the right side of (3.4). For instance, this is the case for the distribution function estimation problem.

In that context we observe independent and identically distributed data pairs $(X_1, Z_1), \dots, (X_n, Z_n)$; we put $Y_{ti} = I(Z_i \le t)$, and $F(t|x) = P(Z \le t|X = x)$; and we take $\widehat{F}(t|x)$ to be the estimator obtained by applying our order-preserving smoother to the datasets $\mathcal{X}_t = \{(X_i, Y_{ti}) : 1 \le i \le n\}$, for $t \in (-\infty, \infty)$. Consider the following assumptions.

C2. The distribution of (X, Z) is compactly supported; the distribution of X is continuous on a compact interval $\mathcal{I}$, has a density that is bounded away from 0 and has two bounded derivatives there, and vanishes off $\mathcal{I}$; the functions $(\partial/\partial t)^j F(t|x)$, for $j = 0, 1, 2$, are bounded uniformly with respect to both $x \in \mathcal{I}$ and $t \in \mathcal{T}$; the bandwidths h_1 used to construct the local linear estimator $\hat{g}_{\mathrm{LL},t}$ from the dataset $\mathcal{X}_t$, and h_2 used to construct the Nadaraya–Watson estimator $\hat{g}_t$ from $\mathcal{X}_t'$, both satisfy $n^\epsilon h_j \to 0$ and $n^{1-\epsilon}h_j \to \infty$ as $n \to \infty$ for some $\epsilon > 0$; the kernels used for either estimator are symmetric, compactly supported, nonnegative, and Hölder continuous on the real line; and the kernel used for $\hat{g}_t$ is unimodal.

Theorem 3.4. Assume conditions C2. Then, with probability 1,

$$\iint \big\{\widehat{F}(t|x) - F(t|x)\big\}^2\,dt\,dx = O\big\{(nh)^{-1} + h^4\big\}.$$

Theorem 3.4 implies a uniform convergence rate for estimators of linear functionals of $F(t|\cdot)$. For example, given a constant $B > 0$, let $\mathcal{C}(B)$ denote the class of differentiable functions ψ satisfying $\sup_t |\psi'(t)| \le B$, and define $\Psi(x|\psi) = E\{\psi(Z)|X = x\}$. Put $\widehat{\Psi}(x|\psi) = \int \psi(t)d_t\widehat{F}(t|x)$. Applying Theorem 3.4 through an integration by parts and an application of Hölder's inequality, we obtain that, with probability 1,

$$\sup_{\psi\in\mathcal{C}(B)} \int \big\{\widehat{\Psi}(x|\psi) - \Psi(x|\psi)\big\}^2\,dx = O\big\{(nh)^{-1} + h^4\big\}.$$

Another case of interest concerns the estimation of conditional quantiles. Assume that, for $0 \le q < r \le 1$, the inverse $F^{-1}(p|x)$ of $F(\cdot|x)$ exists for $p \in [q, r]$ and that $\inf_{p\in[q,r]}\inf_{x\in\mathcal{X}} F'(F^{-1}(p))|x) > 0$. Then, choosing for example as estimates of conditional quantiles

$$\widehat{F}^{-1}(p|x) = \tfrac{1}{2}\Big[\inf_{t\in\mathcal{T}}\big\{t : \widehat{F}(t|x) \ge p\big\} + \sup_{t\in\mathcal{T}}\big\{t : \widehat{F}(t|x) < p\big\}\Big],$$

the result (3.4), applied to conditional distribution functions and combined with bounds for the difference of inverses of two functions, leads to, with probability 1,

$$\sup_{p\in[q,r]}\sup_{x\in\mathcal{X}}\big|\widehat{F}^{-1}(p|x) - F^{-1}(p|x)\big| = O\big\{(nh)^{-1/2}(\log n)^{1/2} + h^2\big\}.$$

4. ILLUSTRATIONS OF ORDER-PRESERVING NONPARAMETRIC REGRESSION

We demonstrate here some simulation- and application-based examples that focus on the case of conditional distribution estimation. Applying the three-step order-preserving nonparametric regression procedure described in Section 2 to data $(X_i, Y_{ti}) = (X_i, I(Z_i \le t))$, for $i = 1, \dots, n$, we first obtain the augmented pseudo-data, (X_i', Y_{ti}'), for $i = 1, \dots, 3n$, and then the bona fide conditional distribution function estimator

$$\widehat{F}(t|x) = \frac{\sum_{i=1}^{3n} K((x - X_i')/h)Y_{ti}'}{\sum_{i=1}^{3n} K((x - X_i')/h)}. \quad (4.1)$$

In the following we choose the kernel function K to be the Bartlett–Parzen–Epanechnikov kernel with support $[-1, 1]$, $K(u) = \frac{3}{4}(1 - u^2)I(-1 \le u \le 1)$. The transition from the initial nonmonotone conditional distribution function estimator $\widehat{F}_{\mathrm{LL}}(\cdot|b)$ to the monotonized version $\widehat{F}(t|b)$ is illustrated in

Figure 1. Generation of the pseudo-data and the final estimate (4.1) is depicted in Figures 2 and 3. If one desires an estimate of the conditional distribution function that is smooth in t, one could use (4.1) as a starting point and then integrate the conditional density kernel estimator $\hat{f}(t|x) = \int \tilde{h}^{-1}\tilde{K}(\tilde{h}^{-1}(t-u))d\widehat{F}(u|x)$ (constructed with kernel $\tilde{K}$ and bandwidth $\tilde{h}$) to obtain the smooth estimate $\widetilde{F}(t|x) = \int_{-\infty}^{t} \hat{f}(v|x)dv$. The preceding formula demonstrates that nonparametric estimation of a conditional density requires a bona fide nonnegative conditional empirical measure, which is only guaranteed if an order-preserving procedure is used.

4.1 Application to Changepoint Estimation and Conditional Distribution Estimation Near Changepoints

We illustrate the use of conditional distribution and quantile function estimation near and at endpoints, through an application to the estimation of changepoint locations and of conditional distribution and quantile functions near changepoints. We consider here a small number of isolated changepoint locations θ, at which the map from the domain of the covariate $[a, b]$ to the space of distribution functions, $x \mapsto F(\cdot|x)$, has a discontinuity, whereas at all other covariate values x it is continuous. Continuity is defined with respect to a suitable metric in the space of distribution functions (Carlstein 1988).

Consider a covariate level x_0 in the interior of $[a, b]$, and denote the conditional distributions (pertaining to the L^2 metric) to the left and right of x_0 as $F_-(\cdot|x_0) = \lim_{x\uparrow x_0} F(\cdot|x)$ and $F_+(\cdot|x_0) = \lim_{x\downarrow x_0} F(\cdot|x)$, respectively. With $D(F_-(\cdot|x_0), F_+(\cdot|x_0)) = \int\{F_+(t|x_0) - F_-(t|x_0)\}^2 dt$, one has $D(F_-(\cdot|x_0), F_+(\cdot|x_0)) = 0$ if the mapping $x \mapsto F(\cdot|x)$ is continuous at $x = x_0$, and $D(F_-(\cdot|x_0), F_+(\cdot|x_0)) > 0$ if a jump occurs at x_0.

We may estimate $F_\pm(\cdot|x_0)$ using the order-preserving estimators $\widehat{F}_\pm(\cdot|x_0)$, by using only the data falling into $[a, x_0]$ when computing $\widehat{F}_-(\cdot|x_0)$, with reflection occurring at the endpoints a and x_0; and analogously for $\widehat{F}_+(\cdot|x_0)$. In both cases, x_0 plays the role of an endpoint.

Using the changepoint detection function $\Delta(\theta) = \int\{\widehat{F}_+(t|\theta) - \widehat{F}_-(t|\theta)\}^2 dt$, the corresponding changepoint location estimate is $\hat{\theta} = \arg\sup_\theta \Delta(\theta)$. If more than one changepoint is to be estimated, this process is simply repeated, by removing the previously estimated locations plus appropriate neighborhoods around them from the set of potential changepoint locations over which a maximum of $\Delta(\cdot)$ is sought.

Once the estimated changepoint location $\hat{\theta}$ has been determined, we set $x_0 = \hat{\theta}$ and obtain order-preserving distribution function estimates using only data where the covariate values fall into the intervals $[a, \hat{\theta}]$ on the left of the estimated changepoint or into the intervals $[\hat{\theta}, b]$ on the right of the estimated changepoint, following exactly the same procedures as for an assumed changepoint at x_0.

4.2 Simulated Example for Change in Variance

We illustrate these procedures first with a simulated dataset. Here

$$Z_i = g(X_i) + e_i, \quad i = 1, \ldots, n,\ X_i \sim U(0,1),\ e_i \sim N(0, \sigma^2),$$

and the e_i's and X_i's are totally independent. Choosing $n = 1,000$, we assume that a change in the variance occurs at $x = .7$, with $g(x) = e^x$ and $\mathrm{var}(e|x) = \sigma^2 = .2, x < .7$, whereas $\mathrm{var}(e|x) = \sigma^2 = 1, x > .7$. The bandwidth was chosen as $h = .2$.

We obtain the changepoint detection function $\Delta(\cdot)$ as shown in Figure 4. A single changepoint location emerges as the

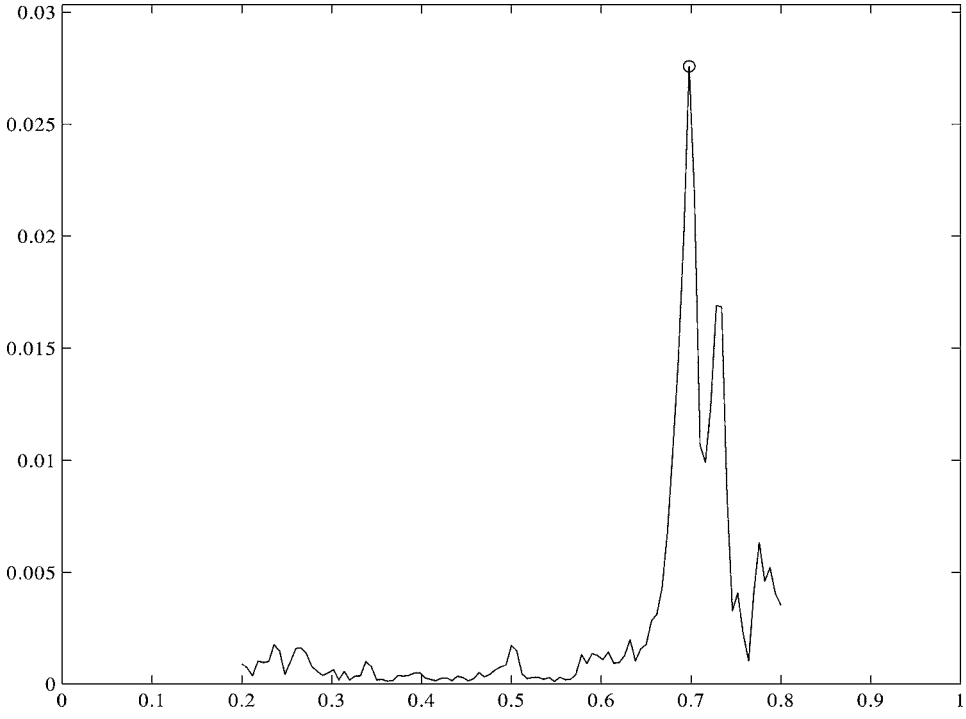

Figure 4. Changepoint Detection Function $\Delta(\cdot)$ for Simulated Variance Change Data ($n = 1{,}000$, h = .2, fixed design case). The peak selected for the changepoint estimate is highlighted.

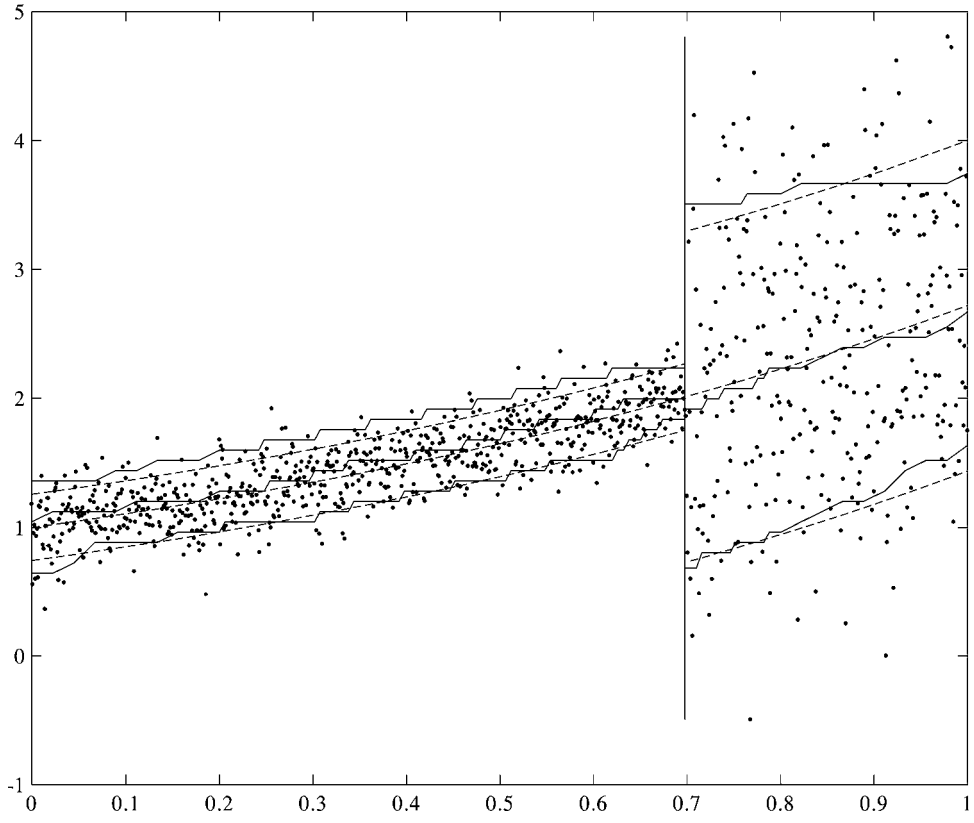

Figure 5. Estimated Conditional Quantile Functions for the Variance Change Data ($h = .2$, n = 1,000). Indicated are the .1, .5, and .9 estimated (solid) and true (dashed) quantile curves. Estimated curves are adapted to the estimated changepoint location.

maximizing argument, and the conditional distribution function estimates adapted to this estimated changepoint location are indicated by the .1, .5, and .9 estimated quantiles in Figure 5. The change in variance is very clearly reflected in these estimates. Figure 6 allows comparison of the two conditional distribution function estimates $\widehat{F}_-(\cdot|\hat{\theta})$ and $\widehat{F}_+(\cdot|\hat{\theta})$ right at the endpoint $\hat{\theta}$, with the corresponding normal underlying distribution functions $F_-(\cdot|\theta)$ and $F_+(\cdot|\theta)$. Agreement is seen to be quite good, confirming that order-preserving estimation of conditional distribution functions performs well at both changepoints and endpoints.

Note that the characteristic of the distribution function that is subject to a sudden change is unknown; that is, it is not assumed to be known in this example that it is a jump in the variance. Commonly used changepoint detection methods based on differences between regular one-sided local linear fits are focusing on mean changes and will not detect this change, because the mean continues to change smoothly across the point of discontinuity in the variance. One-sided bona fide distribution function estimates at endpoints then allow one to define a general detection function Δ based on a suitable distance measure between left- and right-sided distribution function estimates.

4.3 Application to Changepoints in DNA Sequences

The analysis of the frequencies of basepairs in DNA sequences has been studied by many authors; see, for example, Braun and Müller (1998), Braun, Braun, and Müller (2000), Chechetkin and Lobzin (1998), and Liö, Politi, Ruffo, and Buiatti (1996) for biological relevance, methodology, and further references. We use the sequence of *Saccharomyces cerevisiae III*, a chromosome of brewer's yeast, to illustrate order-preserving conditional distribution function estimation in the presence of changepoints.

The data consist of $n = 526$ relative frequencies (obtained by binning) of the occurrence of guanine and cytosine (G+C) as a proportion of all bases (A, C, G, and T). These data are available from Genbank (*http://www.ncbi.nlm.nih.gov/Genbank/*). We assume two changepoints, based on previous analyses of these data. One could extend existing inference procedures as in Dümbgen (1991) or Braun et al. (2000) for the existence and number of changepoints to the more complex situation in this application. Our analysis proceeds by first locating the changepoints with the methods described previously and then constructing the order-preserving conditional distribution function estimates, adapted to the two estimated changepoints as shown via the quantile estimates in Figure 7. Interesting mean and variance patterns become visible, which may motivate further, detailed analyses.

5. THEORETICAL ARGUMENTS

5.1 Proof of Theorem 2.1

Recall that $\mathcal{X}'_t = \{(X'_i, Y'_{ti}) : 1 \le i \le 3n\}$, where $X_1 \le \cdots \le X_{3n}$. We must establish that if $s \le t$, then, for $x \in \mathcal{I}$ and $h \le \frac{1}{2}(b-a)$,

$$\hat{g}_t(x) - \hat{g}_s(x) = \sum_{i=1}^{3n} w_i(x)(Y'_{ti} - Y'_{si}) \ge 0. \qquad (5.1)$$

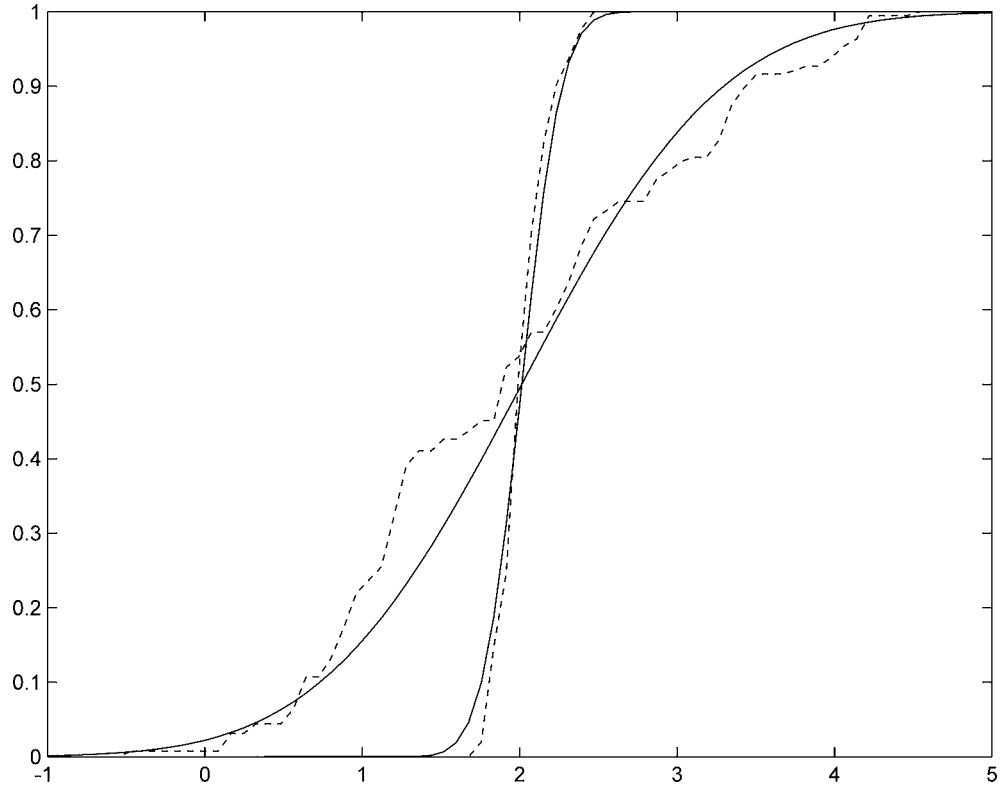

Figure 6. Estimated Conditional Distribution Functions $\widehat{F}_{-}(t|\hat{\theta})$ (steep dashed function) and $\widehat{F}_{+}(t|\hat{\theta})$ (less steep dashed function) and the True Gaussian Conditional Distribution Functions $F_{-}(t|\theta)$ (steep solid function) and $F_{+}(t|\theta)$ (less steep solid function). The distribution function estimates use only data on left or right side of the estimated changepoint.

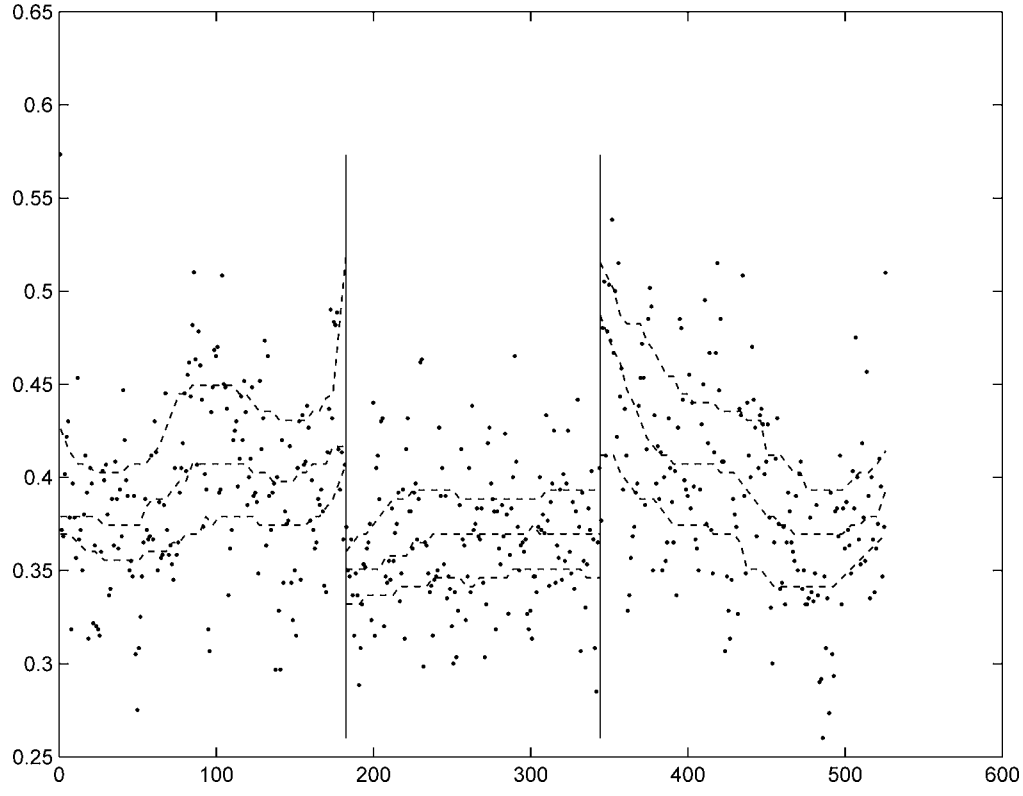

Figure 7. Estimated Order-Preserving Conditional Quantile Functions ($h = 50$ kbp) for the S. cerevisiae III DNA Sequence, Indicated by Estimated .25, .5, and .75 Quantile Curves and Adapted to the Two Estimated Changepoint Locations.

If $n+1 \le i \le 2n$, then (X'_i, Y'_{ti}) is one of the original data, and so, by assumption (see Sec. 2), $Y'_{ti} - Y'_{si} \ge 0$, implying that the corresponding contribution to the series in (5.1) is nonnegative. If $1 \le i \le n$, then either $Y'_{ti} - Y'_{si} \ge 0$, in which case the contribution to (5.1) is nonnegative, or $Y'_{ti} - Y'_{si} < 0$. In the latter case the manner of generation of the pseudo-data implies that there exists an index j, between $n+1$ and $2n$, such that X'_j is the same distance to the right of $x = a$ as X'_i is to the left. Because $\tilde{g}_u(a)$ is nondecreasing as a function of u, $|Y'_{ti} - Y'_{si}| \le Y'_{tj} - Y'_{sj}$. In view of (2.3), $w_i(x) \le w_j(x)$, and so the net contribution of the ith and jth terms to the series in (5.1) is nonnegative. The case $2n+1 \le i \le 3n$ may be treated similarly. The assumption that $h \le \frac{1}{2}(b-a)$ implies that no real data pair needs to be combined in this manner with more than one pseudo-data pair. Hence, for each negative summand in the series in (5.1), there is a positive summand that is not less than the absolute value of the negative summand; and no positive summand has to be combined with two or more negative summands in this way. This establishes the inequality at (5.1).

5.2 Proof of Theorem 3.1

Property (3.2) implies that $\sum_i w_i(a) = 1 + o_p(h)$. Recall that positivity means $\hat{g}(x) \ge 0$ whenever each $Y_i \ge 0$, and so implies that each $w_i(x) \ge 0$. From these results, Taylor expansion, and both parts of (3.1), we may prove that, when $g'(a) \ge 0$,

$$\begin{aligned}
&E\{\hat{g}(a) | X_1, \ldots, X_n\} \\
&\quad = \sum_{i=1}^n w_i(a) g(X_i) \\
&\quad = \sum_{i=1}^n w_i(a) \{g(a) + (X_i - a) g'(a)\} + o_p(h) \\
&\quad = g(a) + g'(a) \sum_{i=1}^n w_i(a)(X_i - a) + o_p(h) \\
&\quad \ge g(a) + g'(a) \sum_{i : |X_i - a| \ge C_2 h} w_i(a)(X_i - a) + o_p(h) \\
&\quad \ge g(a) + C_2 C_3 h g'(a) + o_p(h),
\end{aligned}$$

which implies the theorem. The case $g'(a) < 0$ may be treated similarly.

5.3 Proof of Theorem 3.2

Without loss of generality, the design interval is $[0, b]$. We shall prove that, with probability tending to 1, there exist $s < t$ such that $\widehat{F}(t|0) - \widehat{F}(s|0) < 0$. This result, the continuity of $\widehat{F}(t|x)$ as a function of x, and the symmetry of behavior at either end of $[0, b]$ imply Theorem 3.2.

Note first that it is possible to choose $\xi > 0$ such that $K(\xi) > 0$ and

$$A(\xi) \equiv \int_0^\infty y^2 K(y) dy - \xi \int_0^\infty y K(y) dy < 0.$$

Let $\delta > 0$ be so small that $K \ge C > 0$ on the interval $[\xi, \xi + \delta]$. Because $nh \to \infty$, with probability tending to 1 as $n \to \infty$, there is at least one $X_i \in [\xi h, (\xi + \delta)h]$. Because $F(\cdot|0)$ is nonsingular, for this i we may choose $s < t$ such that $Z_i \in (s, t)$ and no other Z_j lies there. (We suppress the dependence of s and t on i.) Then

$$\widehat{F}(t|0) - \widehat{F}(s|0) = \frac{S_2 - (X_i/h) S_1}{S_2 S_0 - S_1^2} K\left(\frac{X_i}{h}\right),$$

where $S_k = (nh)^{-1} \sum_i (X_i/h)^k K(X_i/h)$. Now $S_2 - (X_i/h) S_1 = A(X_i/h) f(0) + o_p(1)$, and as $n \to \infty$, $S_2 S_0 - S_1^2$ converges in probability to

$$f(0)^2 \left[\frac{1}{2} \int_0^\infty y^2 K(y) dy - \left\{ \int_0^\infty y K(y) dy \right\}^2 \right] > 0,$$

where f denotes the marginal density of X. Hence, in view of our choice of X_i, the probability that $\widehat{F}(t|0) - \widehat{F}(s|0) < 0$ converges to 1 as $n \to \infty$, as had to be shown.

5.4 Outline Proof of Theorem 3.3

Define $\mu_{\mathrm{LL},t}(x) = E\{\hat{g}_{\mathrm{LL},t}(x) | X_1, \ldots, X_n\}$. Using methods based on moderate deviations of sums of independent random variables and, in particular, employing the assumption (3.3) of a finite moment generating function, we may show that there exist $C_1, C_2 > 0$ such that, for all sufficiently large $C_3 > 0$,

$$\begin{aligned}
&\sup_{t \in \mathcal{T}} \max_{x = a, b} P\left\{ \left| \hat{g}_{\mathrm{LL},t}(x) - \mu_{\mathrm{LL},t}(x) \right| > C_3 (nh)^{-1/2} (\log n)^{1/2} \right\} \\
&\qquad = O\left(n^{-C_1 C_3}\right), \qquad (5.2) \\
&\sup_{t \in \mathcal{T}} \max_{x = a, b} P\left\{ |\mu_{\mathrm{LL},t}(x) - g_t(x)| > C_2 h^2 \right\} = O\left(n^{-C_3}\right).
\end{aligned}$$

From these results, the fact that $\mathcal{T}$ has only $O(n^{C_4})$ elements for some $C_4 > 0$, and the Borel–Cantelli lemma, we may prove that, with probability 1,

$$\sup_{t \in \mathcal{T}} \max_{x = a, b} \left| \hat{g}_{\mathrm{LL},t}(x) - g_t(x) \right| = O\left\{ (nh)^{-1/2} (\log n)^{1/2} + h^2 \right\}.$$

The latter property and the definition of $\tilde{g}_t$ imply that, with

$$\xi_t(x) = \tilde{g}_t(x)$$

and

$$\eta = (nh)^{-1/2} (\log n)^{1/2} + h^2, \qquad (5.3)$$

we have, with probability 1,

$$\sup_{t \in \mathcal{T}} \max_{x = a, b} \left| \xi_t(x) - g_t(x) \right| = O(\eta). \qquad (5.4)$$

Assume the kernel used to construct the Nadaraya–Watson estimator $\hat{g}_t$ is supported on $[-C, C]$. Then $\hat{g}_t$ restricted to $[a + Ch, b - Ch]$ does not involve any of the pseudo-data. Define $\mu_t(x) = E\{\hat{g}_t(x) | X_1, \ldots, X_n\}$. Using the arguments leading to (5.2), we may prove that, for constants $C_1, C_2 > 0$, we have, for all sufficiently large $C_3 > 0$,

$$\begin{aligned}
&\sup_{t \in \mathcal{T}} \sup_{a + Ch \le x \le b - Ch} P\left\{ \left| \hat{g}_t(x) - \mu_t(x) \right| > C_3 (nh)^{-1/2} (\log n)^{1/2} \right\} \\
&\quad = O\left(n^{-C_1 C_3}\right), \\
&\sup_{t \in \mathcal{T}} \max_{x = a, b} P\left\{ \left| \mu_t(x) - g_t(x) \right| > C_2 h^2 \right\} = O\left(n^{-C_3}\right).
\end{aligned}$$

From this result, the fact that the kernel used to construct $\hat{g}_t$ is Hölder continuous, the property that $\mathcal{T}$ has only polynomially

many elements, and the Borel–Cantelli lemma, we may show that, with probability 1,

$$\sup_{t\in\mathcal{T}} \max_{a+Ch\le x\le b-Ch} \left|\hat{g}_t(x) - g_t(x)\right| = O(\eta). \tag{5.5}$$

Now consider a generalized form of the estimator $\hat{g}_t(x)$ in which, rather than generating pseudo-data by reflecting the real data through points $(a, \tilde{g}_t(a))$ and $(b, \tilde{g}_t(b))$, we reflect through $(a, \xi_t(a))$ and $(b, \xi_t(b))$, where $\xi_t(x)$ is nondecreasing in t for $x = a, b$. Let $\hat{g}_{\xi_t,t}$ denote the resulting version of $\hat{g}_t$. Using the linearity of Nadaraya–Watson estimators, we may, for $|x-a| \le Ch$, write $\hat{g}_{\xi_t,t}(x) = \hat{g}_{g_t,t}(x) + \{\xi_t(a) - g_t(a)\}\zeta_t(x)$, where the function ζ_t does not depend on the choice of ξ_t. Arguments leading to (5.5) may be employed to prove that, with probability 1,

$$\sup_{t\in\mathcal{T}} \sup_{|x-a|\le Ch} \left|\hat{g}_{g_t,t}(x) - g_t(x)\right| = O(\eta),$$

$$\sup_{t\in\mathcal{T}} \sup_{|x-a|\le Ch} \left|\zeta_t(x)\right| = O(1).$$

Making the choice of ξ_t at (5.3) and noting that (5.4) holds for this selection, we deduce that

$$\sup_{t\in\mathcal{T}} \max_{|x-a|\le Ch} \left|\hat{g}_t(x) - g_t(x)\right| = O(\eta). \tag{5.6}$$

A similar property holds at the other boundary. Combining that result with (5.5) and (5.6), we deduce Theorem 3.3.

Derivation of Theorem 3.4 may similarly be based on application of the Borel–Cantelli lemma, this time using Markov's inequality and moment bounds, as well as techniques from the proof of Theorem 3.3, to prove that, for some $C_1 > 0$ and all $C_2 > 0$,

$$P\left[\iint \left\{\widehat{F}(t|x) - F(t|x)\right\}^2 dt\,dx > C_1\left\{(nh)^{-1} + h^4\right\}\right] = O\left(n^{-C_2}\right).$$

6. CONCLUDING REMARKS

The proposed order-preserving nonparametric regression algorithm provides a first solution to a problem that was previously not tractable, namely, to ensure that nonparametric regression function estimators respect order relationships within the responses, in the interior of the range of the covariate as well as near or at endpoints of the range. We illustrate the importance of the problem and the efficacy of the proposed solution in the case of conditional distribution function estimation. Order-preserving estimation is here a prerequisite for defining bona fide conditional distribution function estimates.

The problem of crossing quantile estimators has been noted and addressed before in linear regression models (He 1997). A comprehensive solution as that given here has not previously been provided. Further relevant applications include the construction of prediction intervals for new observations that are made near endpoints as well as the changepoint problem.

We have shown that no linear method, and this includes practically all commonly used nonparametric regression methods, has the property of being order preserving on the whole domain of the predictor variable. The proposed nonlinear procedure works well and has been shown to possess attractive asymptotic properties. Topics of interest for future research are the development of other order-preserving nonparametric regression methods and an investigation of asymptotic distributions.

[Received May 2002. Revised May 2003.]

REFERENCES

Bhattacharya, P. K., and Gangopadhyay, A. K. (1990), "Kernel and Nearest-Neighbor Estimation of a Conditional Quantile," *Annals of Statistics*, 18, 1400–1415.

Braun, J. V., Braun, R. K., and Müller, H. G. (2000), "Multiple Change-Point Fitting via Quasi-likelihood, With Application to DNA Sequence Segmentation," *Biometrika*, 87, 301–314.

Braun, J. V., and Müller, H. G. (1998), "Statistical Methods for DNA Segmentation," *Statistical Science*, 13, 142–162.

Carlstein, E. (1988), "Nonparametric Change-Point Estimation," *Annals of Statistics*, 16, 188–197.

Chechetkin, V. R., and Lobzin, V. V. (1998), "Study of Correlations in Segmented DNA Sequences: Applications to Structural Coupling Between Exons and Introns," *Journal of Theoretical Biology*, 190, 69–83.

Choi, E., Hall, P., and Rousson, V. (2000), "Data Sharpening for Bias Reduction in Nonparametric Regression," *Annals of Statistics*, 28, 1339–1355.

Cline, D. B. H., and Hart, J. D. (1991), "Kernel Estimation of Densities With Discontinuities or Discontinuous Derivative," *Statistics*, 22, 69–84.

Dümbgen, L. (1991), "The Asymptotic Behavior of Some Nonparametric Change-Point Estimators," *Annals of Statistics*, 19, 1471–1495.

Fan, J. Q. (1992), "Design-Adaptive Nonparametric Regression," *Journal of the American Statistical Association*, 87, 998–1004.

Fan, J. Q., and Gijbels, I. (1992), "Variable Bandwidth and Local Linear Regression Smoothers," *Annals of Statistics*, 20, 2008–2036.

Hall, P., and Presnell, B. (1999), "Intentionally Biased Bootstrap Methods," *Journal of the Royal Statistical Society*, Ser. B, 61, 143–158.

Hall, P., and Wehrly, T. E. (1991), "A Geometrical Method for Removing Edge Effects From Kernel-Type Nonparametric Regression Estimators," *Journal of the American Statistical Association*, 86, 665–672.

Hall, P., Wolff, R. C. L., and Yao, Q. W. (1999), "Methods for Estimating a Conditional Distribution Function," *Journal of the American Statistical Association*, 94, 152–163.

He, X. M. (1997), "Quantile Curves Without Crossing," *The American Statistician*, 51, 186–192.

Liö, P., Politi, A., Ruffo, S., and Buiatti, M. (1996), "Analysis of Genomic Patchiness of Haemophilus Influenzae and Saccharomyces Chromosomes," *Journal of Theoretical Biology*, 183, 455–469.

Mammen, E., and Marron, J. S. (1997), "Mass Centred Kernel Smoothers," *Biometrika*, 84, 765–777.

Müller, H.-G. (1997), "Density Adjusted Kernel Smoothers for Random Design Nonparametric regression," *Statistics and Probability Letters*, 36, 161–172.

Müller, H.-G., and Song, K.-S. (1993), "Identity Reproducing Multivariate Nonparametric Regression," *Journal of Multivariate Analysis*, 46, 237–253.

Peracchi, F. (2002), "On Estimating Conditional Quantiles and Distribution Functions," *Computational Statistics Data Analysis*, 38, 433–447.

Schuster, E. F. (1985), "Incorporating Support Constraints into Nonparametric Estimation of Densities," *Communications in Statistics, Part A—Theory and Methods*, 14, 1123–1136.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, London: Chapman & Hall.

Simonoff, J. S. (1996), *Smoothing Methods in Statistics*, New York: Springer-Verlag.

Wand, M. P., and Jones, M. C. (1995), *Kernel Smoothing*, London: Springer-Verlag.

Yu, K. M., and Jones, M. C. (1998), "Local Linear Quantile Regression," *Journal of the American Statistical Association*, 93, 228–237.